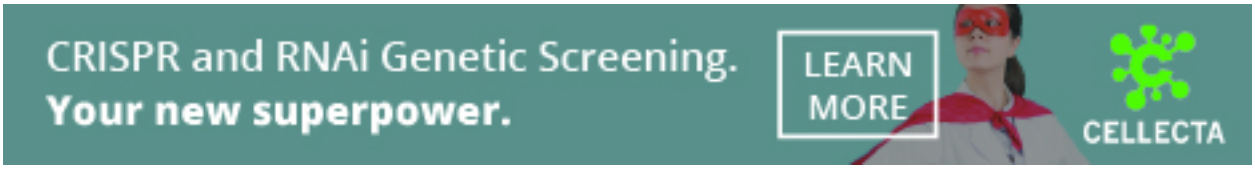
Inferring Alternative Splicing Patterns in Mouse from a Full-Length cDNA Library and Microarray Data

Hiromi Kochiwa, Ryosuke Suzuki, Takanori Washio, et al.

Genome Res. 2002 12: 1286-1293
Access the most recent version at doi:10.1101/gr.220302



Inferring Alternative Splicing Patterns in Mouse from a Full-Length cDNA Library and Microarray Data

Hiromi Kochiwa,[1,3] Ryosuke Suzuki,[1,3] Takanori Washio,[3] Rintaro Saito,[4] The RIKEN Genome Exploration Research Group Phase II Team,[4,5] Hidemasa Bono,[4] Piero Carninci,[4] Yasushi Okazaki,[4] Rika Miki,[4] Yoshihide Hayashizaki,[4] and Masaru Tomita[2,3,6]

[1]Graduate School of Media and Governance, [2]Department of Environmental Information, and [3]Institute for Advanced Biosciences, Keio University, Fujisawa, Kanagawa 252-8520, Japan; [4]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Although many studies on alternative splicing of specific genes have been reported in the literature, the general mechanism that regulates alternative splicing has not been clearly understood. In this study, we systematically aligned each pair of the 21,076 cDNA sequences of *Mus musculus*, searched for putative alternative splicing patterns, and constructed a list of potential alternative splicing sites. Two cDNAs are suspected to be alternatively spliced and originating from a common gene if they share most of their region with a high degree of sequence homology, but parts of the sequences are very distinctive or deleted in either cDNA. The list contains the following information: (1) tissue, (2) developmental stage, (3) sequences around splice sites, (4) the length of each gapped region, and (5) other comments. The list is available at http://www.bioinfo.sfc.keio.ac.jp/intron. Our results have predicted a number of unreported alternatively spliced genes, some of which are expressed only in a specific tissue or at a specific developmental stage.

Alternative splicing of pre-mRNA plays an important role in the production of diverse mRNAs from individual genes, and it helps increase the functional range of gene products in higher eukaryotes. In many cases, gene expression is tightly regulated at the splicing level by specific mechanisms to provide suitable proteins for a particular tissue or stage (McKeown 1992; Chabot 1996; Wang and Manley 1997). On the other hand, alternative transcripts are generated in the same tissue, especially in brain or muscle, to supply an extensive number of proteins that have distinct functions, contributing to their plasticity (Bernstein et al. 1986; Missler and Sushof 1998). The total number of genes in the human genome is estimated to range from 28,000 to 120,000 (Crollius et al. 2000; Ewing and Green 2000; Liang et al. 2000; Wright et al. 2001), and at least one-third of them might give rise to alternatively spliced transcripts (Mironov et al. 1999; Brett et al. 2000). Although the databases of alternative splicing were established by collecting alternatively spliced genes from annotated databases (Dralyuk et al. 2000; Ji et al. 2001), the number of alternatively spliced genes cataloged in such databases is small compared with the estimated total number of alternatively spliced human genes (Modrek et al. 2001).

Using the approach of single-pass end sequence from randomly selected cDNA clones, >1 million expressed sequence tags (ESTs) have been submitted to publicly available databases (Adams et al. 1991). The accumulation of ESTs contributes not only to the discovery of new genes (Adams et al. 1995) but also to the detection of new alternatively spliced genes. There are several ways to detect alternatively spliced genes, including (1) mapping EST sequences onto the genome sequence (Wolfsberg and Landsman 1997; Modrek et al. 2001), (2) comparing full-length mRNA sequences from annotated databases against the EST database (Brett et al. 2000), and (3) clustering EST sequences (Burke et al. 1998). Although the ESTs are effective material to identify novel candidates of alternatively spliced genes, full-length cDNAs are much more desirable for that purpose because they cover entire coding regions.

In this study, we used 21,076 full-length cDNA clones of *Mus musculus* derived from numerous tissues or developmental stages (The RIKEN Genome Exploration Research Group Phase II and the FANTOM Consortium 2001) to analyze the extent of alternative splicing. Here, we conducted a systematic analysis to extract putative alternative cDNAs by comprehensive, round-robin comparisons among the 21,076 clone sequences and constructed a list of potential alternatively spliced transcripts. After that, we analyzed the expression patterns of clusters using their expression profile (Miki et al. 2001) and adopted the clusters whose cDNAs showed a tendency to express in a specific tissue or developmental stage. It has been reported that 69 out of 1600 rat genes were detected

[5]The RIKEN Genome Exploration Research Group Phase II Team: Jun Kawai, Akira Shinagawa, Kazuhiro Shibata, Masayasu Yoshino, Masayoshi Itoh, Yoshiyuki Ishii, Takahiro Arakawa, Ayako Hara, Yoshifumi Fukunishi, Hideaki Konno, Jun Adachi, Shiro Fukuda, Katsunori Aizawa, Izawa Masaki, Katsuo Nishi, Hidenori Kiyosawa, Shinji Kondo, Itaru Yamanaka, and Tetsuya Saito.

[6]Corresponding author.

E-MAIL mt@sfc.keio.ac.jp; FAX 81 (466) 47-5099.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.220302. Article published online before print in July 2002.

12:1286–1293 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

as alternatively spliced genes based on expression data (Hu et al. 2001). Our analysis used a putative alternative splicing data set and an enormous microarray data set.

The use of this method is significant not only because it allowed alternatively spliced genes to be identified but also because it can be limited to the specific condition of alternative splicing and reduce experimental work. This method may be a model of transcriptome analysis of alternative splicing.

RESULTS

Overview of the Clusters Predicted as Alternatively Spliced Genes

The data set of alternatively spliced cDNAs was constructed from a library of 21,076 cDNAs as described in the previous section. The data set consists of 415 clusters with a total of 1136 cDNAs. In the data set, potentially alternatively spliced cDNAs are listed with the following information: (1) tissue, (2) developmental stage, (3) sequences around splice sites, (4) the length of each gapped region, and (5) other comments. These cDNAs are available at http://www.bioinfo.sfc.keio.ac.jp/intron. Most clusters have only one gapped region (putative alternatively spliced site), as summarized in Table 1.

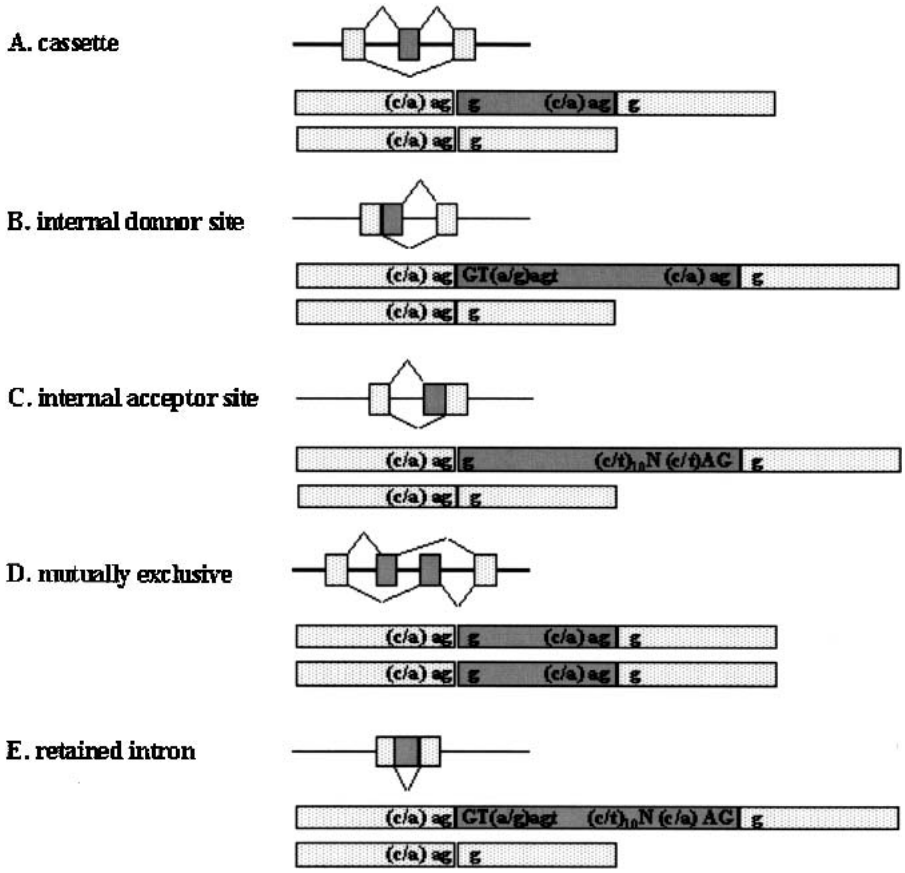


Figure 1 Patterns of alternative splicing. Nucleotide sequences are consensus sequences around the splicing sites (Mount 1982; Padgett et al. 1986).

Various types of alternative splicing patterns have been discussed. Breitbart et al. (1987) suggested five canonical types of alternative splicing (illustrated in Fig. 1): (A) cassette, (B) internal donor site, (C) internal acceptor site, (D) mutually exclusive, and (E) retained intron. We classified the 490 gapped regions of the 415 clusters into one of these five categories according to the criteria defined below. For the sake of classification, we consider nucleotide sequences around the splicing sites (Mount 1982; Padgett et al. 1986) 5′-(a/c)ag|GT(a/g)agt and $(c/t)_{10}$N(c/t)AG|g-3′. These consensus nucleotides are reflected in Figure 1. For each gapped region to be classified into one of the five categories, the nucleotides represented by capital letters are compulsory, and the nucleotides represented by lower-case letters are preferred. More precisely, we used the following criteria: (A) cassette: GT or AG; (B) internal donor site: GT required, and at least four of the seven preferred nucleotides of donor site; (C) internal acceptor site: AG required, and at least 8 of the 13 preferred nucleotides of acceptor site; and (E) retained intron: GT—AG required, and at least four of the seven preferred nucleotides of donor site and 8 of the 13 preferred nucleotides of acceptor site. Because category D can be uniquely determined by the pattern of alignment alone, no nucleotides were checked for it. The gapped regions that could not be classified in each category were categorized as Unclassified. The results of this categorization are presented in Table 2. To estimate the tendency of misclassifications, alternative exons of *M. musculus* known in the literature (Stamm et al. 2000) were used as a

Table 1. Clusters by the Number of Gapped Regions

No. of gapped regions	No. of clusters
One gap	346
Two gaps	48
Three gaps	18
More than three gaps	3
Total	415

Table 2. Classification of Potential Sites of Alternative Splicing

Patterns	No. of gapped regions
(A) Cassette	111
(B) Internal donor site	56
(C) Internal acceptor site	134
(D) Mutually exclusive	8
(E) Retained intron	125
Unclassified	56
Total	490

Table 3. **Known Alternative Exons of** *Mus musculus* Were Classified According to the Same Criteria

	Patterns (actual)			
Patterns (predicted)	**(A) cassette**	**(B) internal donor site**	**(C) internal acceptor site**	**(E) retained intron**
(A) Cassette	26	0	0	0
(B) Internal donor site	2	5	0	0
(C) Internal acceptor site	15	0	9	1
(E) Retained intron	3	2	1	7
Unclassified	14	0	0	0
Total	60	7	10	8

sample set and classified according to the same criteria. The result of this classification is represented in Table 3. The majority of the known exons were categorized correctly in accordance with their appropriate splicing patterns, except many (A) cassette exons were classified as (C) internal accepter sites. These misclassifications arise from the fact that exonic consensus sequences in the acceptor site are similar to the intronic consensus sequence AG, making it difficult to predict the form of alternative splicing on the basis of sequence data (Thanraj 2000). From this control study, it can be inferred that a good portion of the 134 gapped regions listed as (C) internal acceptor sites in Table 2 are actually (A) cassettes.

The numbers of spliced and unspliced regions (illustrated in Fig. 2) of putative alternative splicing are summarized in Tables 4 and 5 according to expressed tissue and developmental stage, respectively. No general tendency specific to tissue or specificity of developmental stage was found, indicating that alternative splicing is taking place widely in all tissues and at all developmental stages.

Details of the Several Clusters Predicted as Alternatively Spliced Genes

One of the clusters in category D (mutually exclusive) is homologous (96% identity) to the CHIP protein (Ballinger et al. 1999). The form of this protein is shown in Figure 3. Although the *CHIP* gene has not been reported as an alternatively spliced gene, it is likely that this gene has alternative transcripts.

Figure 4 shows examples of more complicated alternative splicing patterns in which three cDNAs were potentially produced in different forms from a single gene. An open reading frame (ORF) was predicted for each cDNA using the RIKEN DECODER program (Fukunishi and Hayashizaki 2001).

In the case that an alternatively spliced region resides in a predicted ORF, it is likely that the spliced exon increases variation of the protein function. In particular, cDNA Cluster 8 has three splicing patterns, and the second spliced region causes a drastic change of amino acids by a frameshift. Although it is possible that this frameshift is caused by a sequencing error, we think it is not, because the frameshifted region includes a zinc finger motif (Table 6). It could be suggested that the variety of zinc finger motifs in the three translation products contributes to variation in gene regulation by altering their DNA-binding sites.

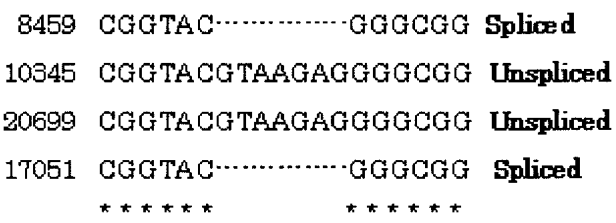


Figure 2 An example of spliced and unspliced regions. Spliced has a gapped region.

Besides this case, frameshifts were identified in cDNA Clusters 63 and 3071, but a motif was not found in these exons. It has been reported that in the integrin β5 subunit of mouse and major protein zero (MPZ) of human, the occurrence of alternative splicing events in the ORF resulted in open-reading frameshifts (Besancon et al. 1999). Thus, two clusters may also have distinct gene functions regulated by frameshifts.

Transcriptome Analysis of Mouse DNA Arrays with Our Data Set

Figures 5 and 6 show the transcriptome analyses of mouse DNA arrays with our putative alternative splicing data set. These clusters each have a prominent splicing pattern in specific tissues or at distinct developmental stages. The level of gene expression is presented as a score of signal intensity between cDNAs.

In Cluster 2204, cDNAs are homologs to prolactin-like peptide. It is known that the prolactin (*PRL*)/growth hormone (*GH*) gene is expressed in the pituitary gland, uterus, or the placenta (Ishibashi and Imai 1999). Our data show that SeqID 4107 is expressed in the placenta but not in the thymus or uterus. On the other hand, SeqID 3784 presents high expression in thymus and uterus. The alternative exon may contribute to the construction of this protein in a particular tissue.

In Cluster 3148, cDNAs are homologs to bisphosphate 3′-nucleotidase (Spiegelberg et al. 1999), which has not been reported to have alternative transcripts. Although the distal start codon may be adopted by both cDNAs, two start codons may be properly used at a specific developmental stage by alternative splicing.

Some alternatively spliced regions are outside of predicted ORFs (Clusters 3082, 3138). The cDNAs of Cluster 3138 are homologs to TIA-1 cytotoxic granule-associated RNA-binding protein-like 1. This gene is expressed in the cells fated to be brain and retina at embryonic days 12.5. Its expression is also found in the lung, kidney, and thymus (Lowin et al. 1996). On the other hand, the gene expression of cDNA Cluster 3082 is likely to be regulated according to the skin developmental stage. The cDNAs of this cluster are homologs to 28S ribosomal protein S17 (Gantt and Thompson 1990). It has been reported that alternative splicing often occurs in 5′-untranslated regions, resulting in alternative regulation of gene expression (Mironov et al. 1999). Therefore, the alternatively spliced regions may contain regulatory elements.

DISCUSSION

We divided 1136 cDNAs into 415 clusters as putative alternatively spliced transcripts. These cDNAs constitute 7.4% of the

Table 4. The Number of Spliced and Unspliced Regions Listed by Tissues

Tissue	No. of gapped regions	
	spliced	unspliced
Adipose	0	1
Brain	4	9
C. quadrigemina microdissected	0	2
Cecum	4	1
Cerebellum	21	15
Cerebellum microdissected	1	1
Colon	0	4
Corpus striatum microdissected	2	3
ES cell	31	24
Extra testis	1	0
Extra testis microdissected	1	3
Eyeball	2	2
Eyeball microdissected	0	1
Forelimb	3	0
Head	36	35
Heart	5	5
Hippocampus	13	8
Hypothalamus microdissected	1	1
Intestine	1	0
Kidney	21	40
Liver	33	16
Liver microdissected	19	16
Lower body	1	1
Lung	25	7
Lung microdissected	0	5
Mammary gland	1	0
Medulla oblonagata microdissected	5	3
Ovary and uterus	5	2
Pancreas	30	39
Pituitary gland	6	5
Placenta	10	6
Placenta and extraembryonic tissues	4	1
Retina microdissected	3	0
Skin	8	9
Small intestine	26	34
Spinal cord microdissected	1	1
Spleen	0	1
Stomach	28	17
Stomach microdissected	0	1
Testis	123	101
Testis microdissected	1	1
Thymus	2	3
Tongue	60	57
Upper body	2	2
Urinary bladder	1	0
Whole body	188	199
Whole body microdissected	1	1
Total	720	683

Table 5. The Number of Spliced and Unspliced Regions Listed by Developmental Stage

Developmental stage	No. of gapped regions	
	spliced	unspliced
Adult	418	375
Embryo-10	52	58
Embryo-10+	45	38
Embryo-11	30	26
Embryo-12	9	11
Embryo-13	45	34
Embryo-14	0	1
Embryo-14, 17	2	3
Embryo-15	1	1
Embryo-16	1	6
Embryo-17	2	3
Embryo-18	34	58
Embryo-7	1	1
Embryo-8	25	19
ES cell	31	24
Lactation-10	1	0
Neonate-0	12	14
Neonate-10	2	6
Neonate-6	4	3
Pregnant-11	5	2
Total	720	683

15,294 cDNAs (the estimated number of unique sequences). Although it has, indeed, been reported that ~38% of all human genes are produced by alternative splicing (Brett et al. 2000), our number should not be interpreted as the percentage of alternatively spliced genes in general. In the process of constructing the cDNA library, we tried to reduce redundancy by not sequencing cDNAs with the same nucleotide sequence in their 5′- or 3′-untranslated regions (The RIKEN Genome Exploration Research Group Phase II and the FANTOM Consortium 2001). This procedure should have eliminated a large number of alternatively spliced transcripts.

It has been reported that many genes are alternatively spliced at multiple sites (Smith et al. 1989), from which hundreds of alternate transcripts could be produced in theory. One example of this is the lymphocyte homing receptor gene *CD44*, which can generate enormous molecular diversity, >1000 potential isoforms, by including or excluding each of 10 exons in the gene (Screaton et al. 1992; Tolg et al. 1993). In our results, on the other hand, most of the clusters showed potential alternative splicing at only one site (Table 1); it may be that they have many more splicing variants that we have overlooked. To study this possibility, a greater amount of cDNA sequence data from a given gene will be necessary (Regan et al. 2000).

In summary, computational analysis is a powerful means for predicting potential sites of alternative splicing, and we have constructed a list of these sites from the largest available data set of mouse full-length cDNA sequences. Our results have predicted a number of unreported alternatively spliced genes, some of which are expressed only in a specific tissue or at a specific developmental stage.

METHODS

We used a set of 21,076 mouse full-length cDNAs produced by The RIKEN Genome Exploration Research Group Phase II and the FANTOM Consortium (2001). The average length of all the cDNAs was 1257 bp. The number of unique sequences, after eliminating redundant sequences, is presumed to be 15,294. In our work, however, we did not make any attempt

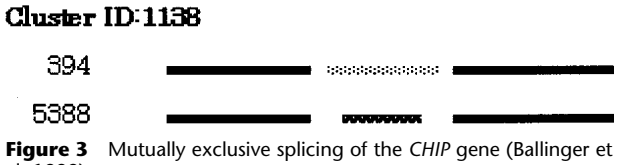


Figure 3 Mutually exclusive splicing of the *CHIP* gene (Ballinger et al. 1999).

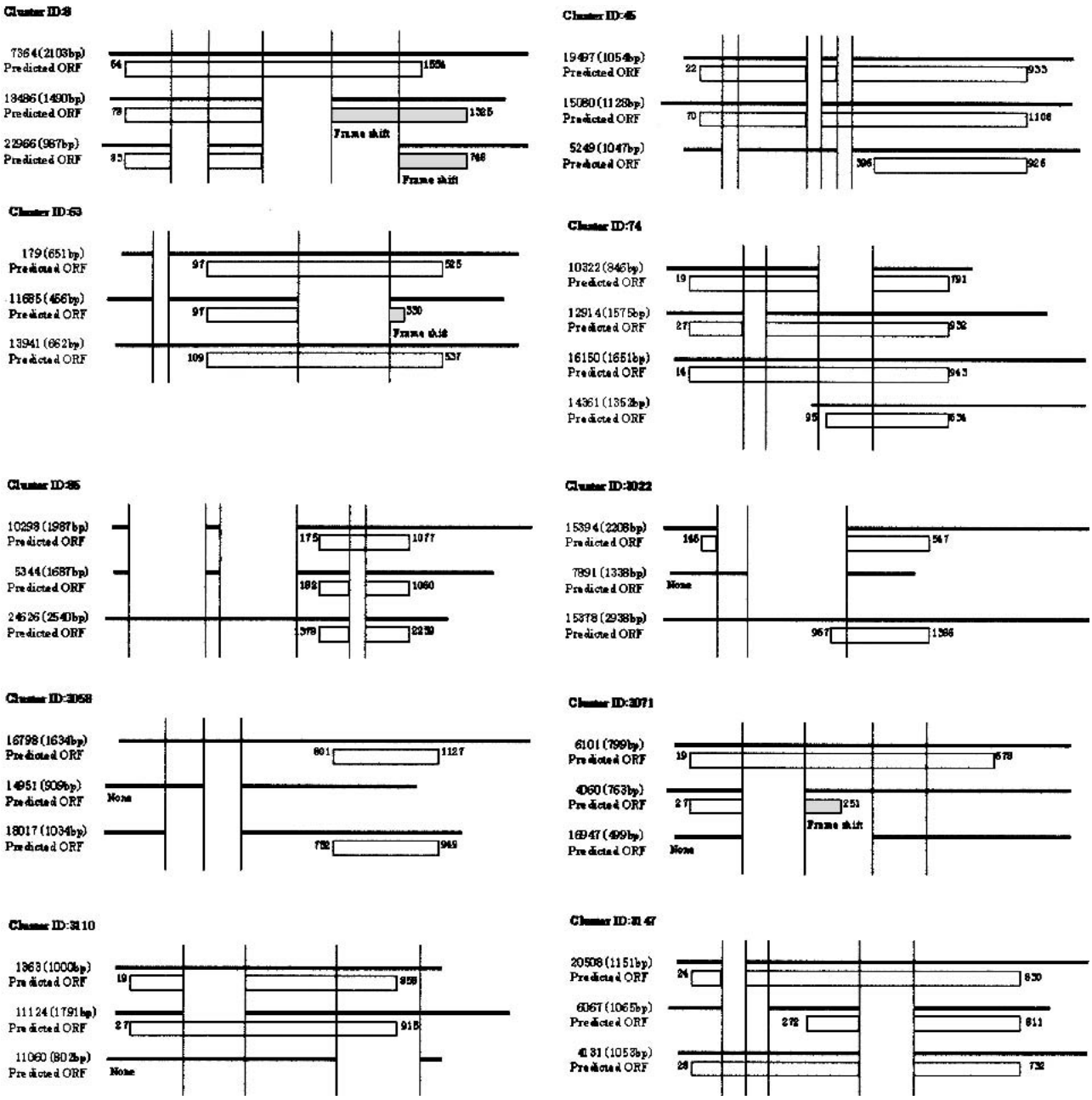


Figure 4 Examples of more complicated alternative splicing patterns in which three cDNAs were potentially produced in different forms from a single gene. Cluster 8: homologs to human PR domain zinc finger protein 5 (Deng et al., unpubl.). Cluster 45: homologs to human mitochondrial carrier homolog 2 (Jang et al., unpubl.). Cluster 63: homologs to human HSPC204 protein (Zhang et al. 2000). Cluster 74: homologs to human HSPC223 protein (Ye et al., unpubl.). Cluster 85: homologs to human heterogeneous nuclear ribonucleoprotein C (Nakagawa et al. 1986). Clusters 3022, 3058, and 3110: no homology found (hypothetical protein). Splice variant of Cluster 3058, no homology found (unclassifiable). Cluster 3147: homologs to *D. melanogaster* brain cDNA clone NMCB-2386 (Osada et al., unpubl.). Cluster 3148: homologs to bisphosphate 3′-nucleotidase (Spiegelberg et al. 1999).

to eliminate redundancy and used all of the 21,076 sequences, in order not to miss any potential alternative transcripts.

First, we conducted a round-robin BLAST search (Altschul et al. 1990) of the 21,076 cDNAs sequences against each other. The cDNA pairs whose BLAST output met the following criteria were extracted from the data set: (1) >95% of nucleotides were identical for >20 consecutive nucleotides; and (2) more than one such matching region in common. After these comprehensive pair-wise comparisons, the cDNA

Table 6. The Result of Motif Analysis in Alternate Exons (Cluster 8)

Seq. ID	Description of motif	Number of motif	*E value*
18486	Zinc finger C2H2 type	11	7.00 *E*-14
7864	Zinc finger C2H2 type	11	1.80 *E*-82
22966	Zinc finger C2H2 type	8	3.20 *E*-66

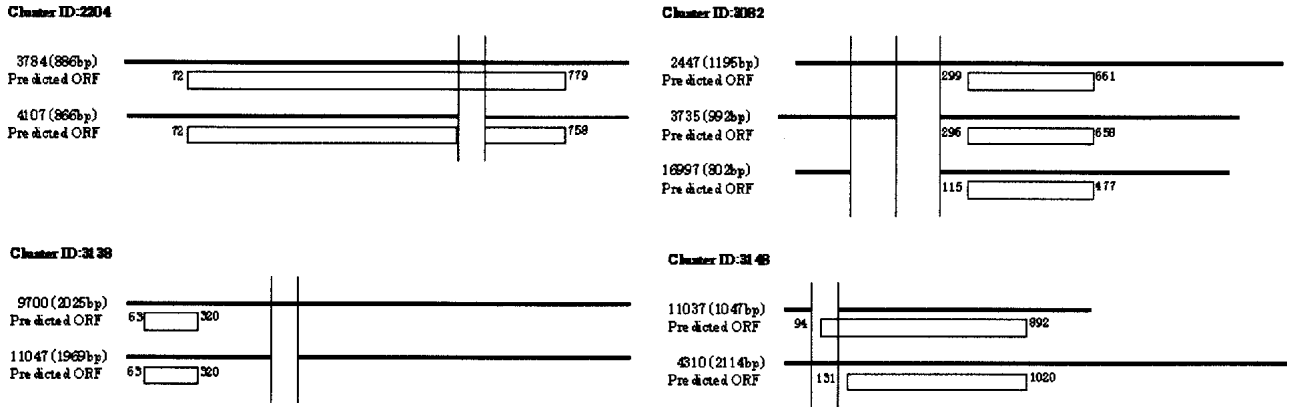


Figure 5 These clusters each have a prominent splicing pattern in specific tissues or at distinct developmental stages. Cluster 2204: homologs to prolactin-like-peptide (Ishibashi and Imai 1999). Cluster 3082: homologs to human HSPC011 and 28S ribosomal protein S17, mitochondrial precursor (Gantt and Thompson 1990). Cluster 3138: homologs to TIA-1 cytotoxic granule-associated RNA-binding protein-like 1 (Lowin et al. 1996). Cluster 3148: homologs to bisphosphate 3′-nucleotidase (Spiegelberg et al. 1999).

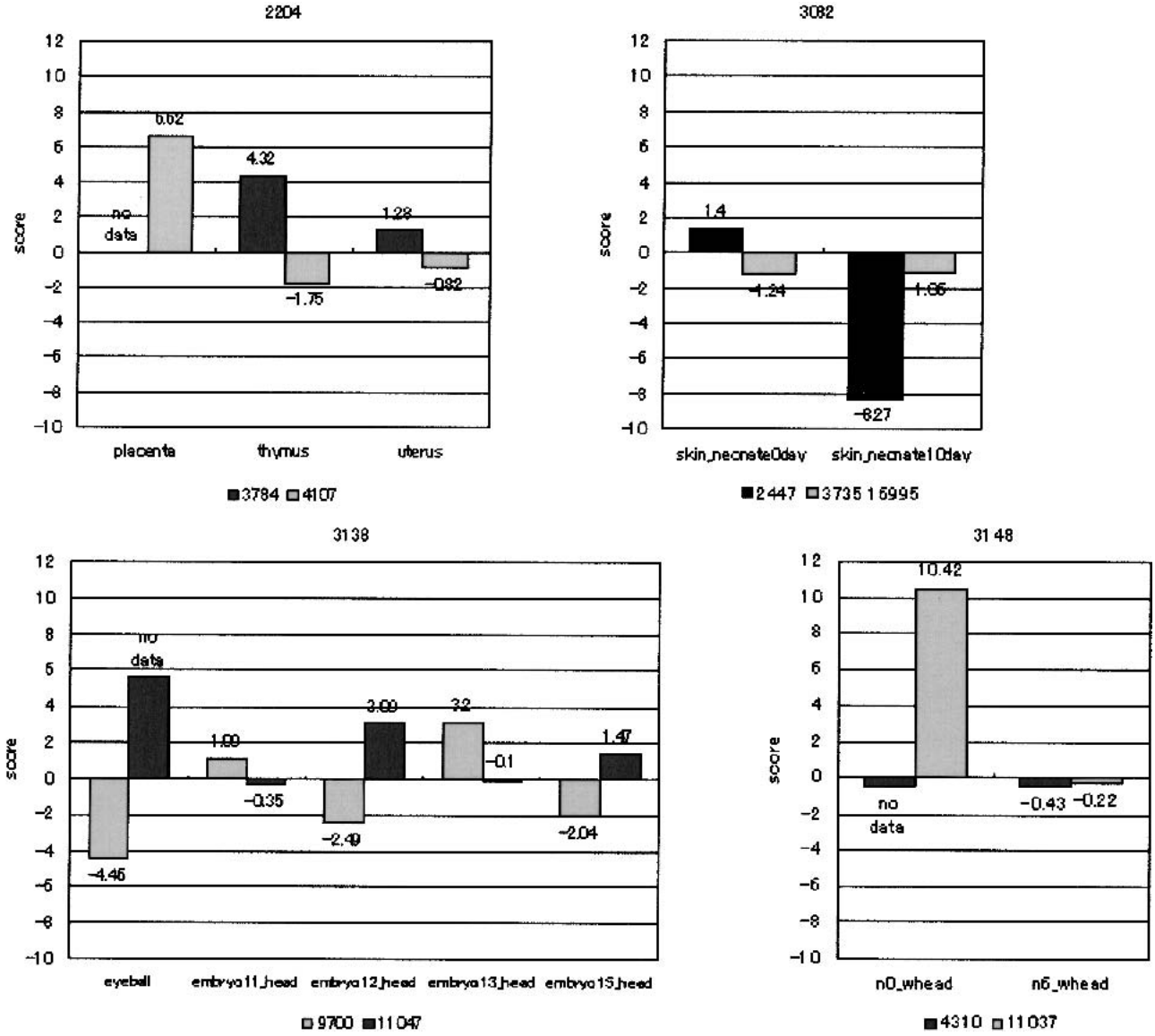


Figure 6 The horizontal axis is the tissue in which the gene expression was observed. The vertical axis is the level of gene expression as a score of signal intensity between cDNAs (log).

pairs were merged into clusters, if one sequence was paired with two or more different sequences.

Next, the sequences of these clusters were aligned using the multiple sequence alignment program CLUSTALW (Thompson et al. 1994). The gap penalty parameter was set to 0 to tolerate large gaps. If the output of alignment shared most of the region with a high degree of sequence homology but parts of the sequences were very distinctive or deleted in either cDNA, the cluster was suspected to be alternatively spliced originating from the common gene. We define such distinctive or deleted regions as gapped regions, and consider them as candidate alternatively spliced exons.

We also used microarray data of expression patterns for 18,816 mouse cDNA sequences (Miki et al. 2001), to extract alternatively spliced genes whose expression pattern is prominent in a specific tissue or at a specific developmental stage. We presented the level of gene expression as a score of signal intensity between cDNAs.

ACKNOWLEDGMENTS

We thank Atsushi Sakurai, Shigeo Fujimori, Koya Mori, Hitomi Itoh, and members of the Tomita laboratory for helpful discussions and suggestions during the course of this work. This study was supported in part by a research grant for the RIKEN Genome Exploration Research Project from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to Y.H. This work was also supported by a research grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project), New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry of Japan (Development of a Technological Infrastructure for Industrial Bioprocesses Project), and Japan Science and Technology Agency.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Ballinger, C.A., Connell, P., Wu, Y., Hu, Z., Thompson, L.J., Yin, L.Y., and Patterson, C. 1999. Identification of CHIP, a novel tetratricopeptide repeat-containing protein that interacts with heat shock proteins and negatively regulates chaperone functions. *Mol. Cell. Biol.* **19:** 4535–4545.

Bernstein, S.I., Hansen, C.J., Becker, K.D., Wassenberg II, D.R., Roche, E.S., Donady, J.J., and Emerson, C.P., Jr. 1986. Alternative RNA splicing generates transcripts encoding a thorax-specific isoform of *Drosophila melanogaster* myosin heavy chain. *Mol. Cell. Biol.* **6:** 2511–2519.

Besancon, R., Prost, A.L., Konecny, L., Latour, P., Petiot, P., Boutrand, L., Kopp, N., Mularoni, A., Chamba, G., and Vandenberghe, A. 1999. Alternative exon 3 splicing of the human major protein zero gene in white blood cells and peripheral nerve tissue. *FEBS Lett.* **457:** 339–342.

Breitbart, R.E., Andreadis, A., and Nadal-Ginard, B. 1987. Alternative splicing: A ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.* **56:** 467–495.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **26:** 83–86.

Burke, J., Wang, H., Hide, W., and Davison, D.B. 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276–290.

Chabot, B. 1996. Directing alternative splicing: Cast and scenarios. *Trends Genet.* **12:** 472–478.

Crollius, R.H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25:** 235–238.

Dralyuk, I., Brudno, M., Gelfand, M.S., Zorn, M., and Dubchak, I. 2000. ASDB: Database of alternatively spliced genes. *Nucleic Acids Res.* **28:** 296–297.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Fukunishi, Y. and Hayashizaki, Y. 2001. Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics* **5:** 81–87.

Gantt, J.S. and Thompson, M.D. 1990. Plant cytosolic ribosomal protein S11 and chloroplast ribosomal protein S17. Their primary structures and evolutionary relationships. *J. Biol. Chem.* **265:** 2763–2767.

Hu, G.K., Madore, S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. 2001. Predicting splice variant from DNA chip expression data. *Genome Res.* **11:** 1237–1245.

Ishibashi, K. and Imai, M. 1999. Identification of four new members of the rat prolactin/growth hormone gene family. *Biochem. Biophys. Res. Commun.* **262:** 575–578.

Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X., and Li, Y. 2001. AsMamDB: An alternative splice database of mammals. *Nucleic Acids Res.* **29:** 260–263.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25:** 239–240.

Lowin, B., French, L., Martinou, J.C., and Tschopp, J. 1996. Expression of the CTL-associated protein TIA-1 during murine embryogenesis. *J. Immunol.* **157:** 1448–1454.

McKeown, M. 1992. Alternative mRNA splicing. *Annu. Rev. Cell Biol.* **8:** 133–155.

Miki, R., Kadota, K., Bono, H., Mizuno, Y., Tomaru, Y., Carninci, P., Itoh, M., Shibata, K., Kawai, J., Konno, H., et al. 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci.* **98:** 2199–2204.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288–1293.

Missler, M. and Sudhof, T.C. 1998. Neurexins: Three genes and 1001 products. *Trends Genet.* **14:** 20–26.

Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29:** 2850–2859.

Mount, S.M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10:** 459–472.

Nakagawa, T.Y., Swanson, M.S., Wold, B.J., and Dreyfuss, G. 1986. Molecular cloning of cDNA for the nuclear ribonucleoprotein particle C proteins: A conserved gene family. *Proc. Natl. Acad. Sci.* **83:** 2007–2011.

Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S., and Sharp, P.A. 1986. Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55:** 1119–1150.

Regan, M.R., Emerick, M.C., and Agnew, W.S. 2000. Full-length single-gene cDNA libraries: Applications in splice variant analysis. *Anal. Biochem.* **286:** 265–276.

The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Screaton, G.R., Bell, M.V., Jackson, D.G., Cornelis, F.B., Gerth, U., and Bell, J.I. 1992. Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. *Proc. Natl. Acad. Sci.* **89:** 12160–12164.

Smith, C.W., Patton, J.G., and Nadal-Ginard, B. 1989. Alternative splicing in the control of gene expression. *Annu. Rev. Genet.* **23:** 527–577.

Spiegelberg, B.D., Xiong, J.P., Smith, J.J., Gu, R.F., and York, J.D. 1999. Cloning and characterization of a mammalian lithium-sensitive bisphosphate 3′-nucleotidase inhibited by inositol 1,4-bisphosphate. *J. Biol. Chem.* **274:** 13619–13628.

Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., and Zhang, M.Q. 2000. An alternative-exon database and its statistical analysis. *DNA Cell Biol.* **19:** 739–756.

Thanraj, T.A. 2000. Positional characterization of false positives from computational prediction of human splice sites. *Nucleic Acids Res.* **28:** 744–754.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Tolg, C., Hofmann, M., Herrlich, P., and Ponta, H. 1993. Splicing choice from ten variant exons establishes CD44 variability. *Nucleic Acids Res.* **21:** 1225–1229.

Wang, J. and Manley, J.L. 1997. Regulation of pre-mRNA splicing in metazoa. *Curr. Opin. Genet. Dev.* **7:** 205–211.

Wright, F.A., Lemon, W.J., Zhao, W.D., Sears, R., Zhuo, D., Wang, J.P., Yang, H, Y., Baer, T., Stredney, D., Spitzner, J., et al. 2001. A draft annotation and overview of the human genome. *Genome Biol.* **2:** RESEARCH0025.1–RESEARCH0025.18.

Wolfsberg, T.G. and Landsman, D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25:** 1626–1632.

Zhang, Q.H., Ye, M., Wu, X.Y., Ren, S.X., Zhao, M., Zhao, C.J., Fu, G., Shen, Y., Fan, H.Y., Lu, G., et al. 2000. Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem progenitor cells. *Genome Res.* **10:** 1546–1560.

WEB SITE REFERENCES

http://www.bioinfo.sfc.keio.ac.jp/intron; a list of alternative splicing patterns.

Received October 26, 2001; accepted in revised form May 17, 2002.